Ontology-Driven Generation of a Federated Schema for GIS

Agustina Buccella[1], Domenico Gendarmi[2], Filippo Lanubile[2], Alejandra Cechich[1], and Attilio Colagrossi[3]

[1] GIISCO Research Group,
Departamento de Ciencias de la Computación,
Universidad Nacional del Comahue, Neuquen,Argentina
{abuccel,acechich}@uncoma.edu.ar
[2] Dipartimento di Informatica,
University of Bari,
Via E. Orabona, 4 - 70125, Bari, Italy
{gendarmi,lanubile}@di.uniba.it
[3] Dipartimento Tutela delle Acque Interne e Marine,
APAT, Via Curtatone, 3 - 00185, Rome, Italy
attilio.colagrossi@apat.it

Abstract. In this work we propose an extension of a Federated System, named Information Broker, developed with the Italian Agency for Environmental Protection and Technical Services (APAT). The main objective of this proposal is to build an integrated system taking into account autonomous, distributed and heterogeneous geographic sources. Our extension is aimed at improving aspects as redundancy, consistency, and scalability by adding semantic interoperability through the use of ontologies and the ISO 19100 standards.

Key words: Geographic Information Systems, Federated Systems, Ontology, ISO 19100 Standards

1 Introduction

The APAT was established in 1999 to carry out scientific and technical activities in the national interest to protect the environment, water resources and soil. Data collected include climatic, hydrometric, cartographic and water pollution measures. Although all the information is owned by the same organization, the huge amount of information is managed by different departments and units. Besides, given the large diversity in syntax and semantic of data, measures are stored into several independent systems, which are based on the most appropriate technology for their data type. All these characteristics have made very hard to share information among the different systems. Thus, the main goal of the APAT Information Broker project is to develop a system to provide a fully and user-transparent integration of the heterogeneous data sources, ensuring at the same time, the existing legacy applications that operates on them will continue operating autonomously, without undergoing any sort of modification.

In previous work [1, 2] we have developed an Information Broker System together with a schema integration process focusing specially on syntactic interoperability. This system is mainly represented by using XML data models for the integration process without storing semantic information. Therefore, the process is made manually, increasing the chance of introducing errors and inconsistences.

In this work, we propose an extension of the schema integration process by adding semantic information through the use of ontologies [3]. Then, the Information Broker System will be implemented as an ontology-driven system in order to share the real common vocabulary contained in the sources. We have focused on ontologies due to the advantages they provide to an integration process – as ontologies are formally described, i.e. by using some logic language such as Description Logic [4], we will be able to perform inferences and check inconsistencies easily.

Our extension is based on previous work on integration of geographic information [5, 6], which focuses on two main aspects: modelling and integrating ontologies. With respect to the former, the ontologies are created towards integration by using a family of the ISO 19100 standards (prepared by ISO Technical Committee 211 (TC211)[4]). Specially ISO 19109 [7], ISO 19110 [8], and ISO 19107 [9] are used in these works. On the other hand, we propose an integration methodology focused on three main phases: *unit*, *integration* and *system*. Each phase takes advantage of the semantic of ontologies and their specific representation. This integration process is mainly based on our work in [5].

This paper is organized as follows: next Section describes the current Information Broker System. Section 3 presents the extension describing its architecture. In Section 4 we discuss some related work. Finally, future work and conclusions are discussed afterwards.

2 The Information Broker System

Distributed and overlapped information in APAT have motivated the construction of a federated system based on hydrological features. As the main goal of the project is to develop a system to provide a fully and user-transparent integration of the sources, in previous work [1, 2] we have introduced and implemented an Information Broker System based on a layered-based architecture. Figure 1 shows this architecture consisting of three main layers, *wrapper*, *federation* and *presentation*.

In the *Wrapper Layer*, Data Access Services (DASs) have been developed to wrap each available data source and to extract the information required on demand. Following, the *Federation Layer* offers a uniform and transparent access to the data stored in data sources through the *Query Processor* and the *Federated Schema Browser* components. The *Query Processor* performs the task of decomposing a global query in a set of local queries and integrating all the obtained results in a single response. The *Federated Schema Browser* provides

[4] http://www.isotc211.org/

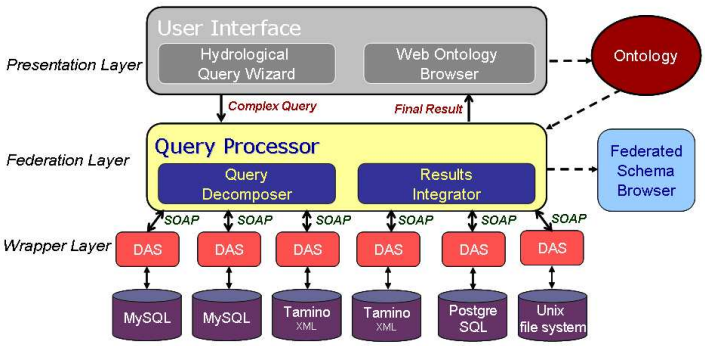

Fig. 1. The Information Broker Architecture

a high-level access to the federated schema, and is used by the query processor to discover the appropriate DAS which, in turn, provides access to a specific concept.

Finally, the *Presentation Layer* represents the communication medium between the broker and the end-users. It consists of two main components: the *User Interface* and the *Ontology*. Two different user-interfaces have been developed: an hydrological query wizard, used to perform global queries and view consequent results in a common web browser; and a web ontology browser, enabling users to navigate through the hydrological concepts (the ontology) within the APAT domain.

We have developed a first prototype of the Information Broker System [2]. This first release is composed of six databases managed by three distinct DBMS, namely MySQL Server, used for collecting real time measures; PostgreSQL server, used for collecting information on water quality; and Tamino XML Server, used for collecting data on extreme hydrological events and hydrography of the territory. Empirical evaluations about the use of this system are still outstanding.

In this paper, we are interested in one of the main processes to build the *Federated Schema* of the federation layer. Next sub-section describes some details of this process.

2.1 Building the Federated Schema

The federated schema is designed to provide a shared vocabulary of the information sources. Based on this vocabulary, we implement the user interface and the query processor components in order to give a global view of the whole system. In this way, the federated schema constitutes the core of the Information Broker system.

A bottom-up process consisting of four steps has been designed taking into account syntactic interoperabilities. Figure 2 shows graphically the components created within each step.

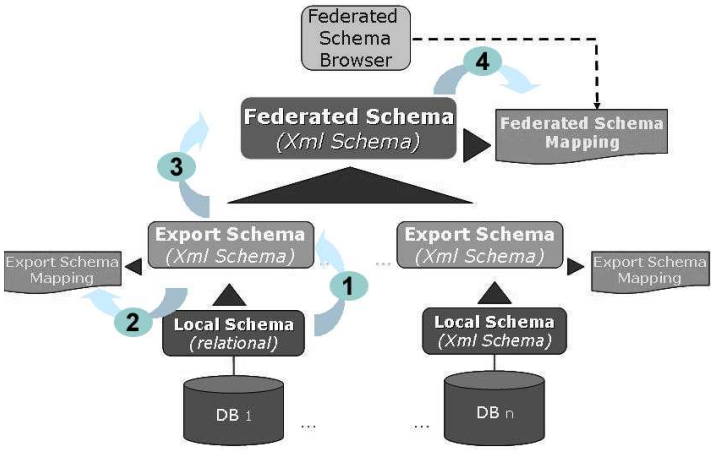

Fig. 2. Schema Integration Process

The first step transforms the local schemas into so-called export schemas, which are expressed in a common data model (CDM) and represented by XML data models. Thus, local schemas of the different databases of the federation converge on a common structure of data.

Then, the second step creates the export-schema mappings, which are XML files manually generated at design time from each export schema. Such files contain the mappings between the local and export schemas; that is, correspondences between low-level data and high-level domain concepts.

Finally, the third step builds the federated schema, which represents the logical model of the virtual database containing all data available within the federation. The federated schema is the result of merging all the export schemas.

During this merging, all possible conflicts must be identified and solved. This is accomplished through two different activities. The *Correspondence Investigation* activity searches for correspondences among the export schemas. The output of this activity is a set of conflicts, grouped in *naming conflicts* and *structural conflicts*. After that, the *Conflict Resolution* activity is carried out reviewing and fixing each conflict.

Once the federated schema has been generated, the last step in the process manually generates the federated-schema mapping file. It consists of an XML file that stores the correspondences between complex concepts and simple concepts distributed in the different export schemas; simple concepts and constraints that characterize them; and simple concepts and services able to retrieve them.

With respect to semantic aspects of the Information Broker System, we add a new component, called *Ontology Schema Mapping*, in order to represent the correspondences between concepts in the domain ontology and queries.

3 An Ontology-Based Extension for Generating a Federated Schema

Although the current Information Broker architecture is well suited for manipulating standard information through XML formatting rules, integration completely depends on users' interpretations and background. As we aforementioned, the task of building the federated schema is completely manual and in the case of large information sources (as we have to consider in this project) it becomes tedious and error-prone. Aspects as modificability and scalability were not taken into account because re-executing the integration process only for some changes on data can take several days.

In this way, the process of building the federated schema becomes difficult to standardize and evolve. Taking into account these two points we propose some changes on the general process of building the federated schema in order to facilitate the use of more suitable processes. The proposed extension is based on previous work [5, 6] in which an architecture and a merging process have been defined.

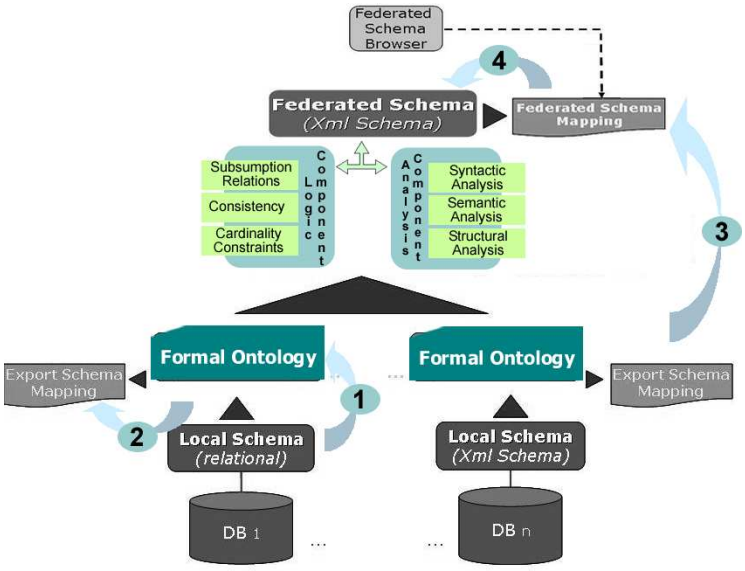

Fig. 3. Changes on the Schema Integration Process

Figure 3 shows the main changes made on the original Information Broker architecture. Like in the original schema (Figure 2), four bottom-up steps are necessary to build the federated schema. However, these steps are very different. The first and second steps, which were in charge of transforming local schemas into export schemas and generating export schema mappings, are now responsible for standardizing the geographic information of sources.

The third and fourth steps, which were in charge of creating the federated schema and its mappings, are now responsible for applying the method for merging ontologies.

In this way, the four steps are combined into two main processes, *enriching local ontologies* and *the merging process* itself. The first process defines the steps to create formal ontologies by applying the ISO 19100 standard for geographic information. Then, the *merging process* implements our merging method. Next two sub-sections provide a brief description of these processes.

3.1 Enriching local ontologies

The use of the ISO 19100 standard gives a new perspective to face integration problems for the interoperability of geographic systems. New ontology modelling techniques of this type of systems should be based on this standard in order to allow integration methods take advantage the benefits they provide.

In our extension, a *top-level ontology* and a *domain ontology* are built based on the information provided by the models of the standard (ISO 19109 and 19107 std.). Gray arrows in the Figure show how the information flows among the models. Thus, the domain ontology is built considering the General Feature Model (GFM) and the Application Schema [7]. The GFM is a meta-model of feature types. It defines the structure for classifying features used then to build the application schema. In the case of the top-level ontology, it is based on the structure of the GFM and the general features of the model being built.

Currently, there are new methodologies proposing the creation of ontologies such as [10, 11], including *Semantic Enrichment* as one of the most important steps. The main goal of this is to reconcile semantic heterogeneity, so it involves adding more semantic information about data. In our work, as both ontologies – top-level and domain – have to be based on the standard before being created, we add a new step in the process named *the enrichment step*. In this step, the components of the ontologies are enriched in their descriptions, through the metaclasses (from GFM) which they are instance of and the schemas on which they are based. In this way, all metaclasses extracted from the GFM and representing information by the application schema are created as abstract classes in the local ontology. Creating an ontology with these characteristics is not a complex task because the information needed with respect to the GFM can be extracted from the Feature Catalogue. Besides, by using an ontology editor as Protégé [5] to model OWL ontologies [12], ISO ontologies from `http://loki.cae.drexel.edu/~wbs/ontology/list.htm` can be imported.

[5] `http://protege.stanford.edu/`

Thus, all the ontologies will have the same structure due to all components are subclassifying the same model. The GFM acts as a top-level ontology classifying the elements of the ontology and making the integration easier. We will discuss this in the next sub-section.

3.2 The Merging Process

The merging process involves the task of merging the geographic sources in order to create a global vocabulary (federated schema) by defining two main components (Figure 3), *logic* and *analysis*. Both processes are used in different parts of the merging process.

This process is composed of three main phases: *unit*, *integration* and *system*. In the *Unit Phase* each system is analyzed separately. The top-level and domain ontologies can be seen as a unique ontology in which generalization / specialization relations are the connectors between them. This ontology will be formally represented by using OWL [12].

Then, once the ontologies are correctly created, a Reasoning System (such as RACER [13]) is applied in order to discover inferences not detected by users. We take advantage of the capability of inferring subsumption relations between classes and properties in the schema (TBox). That is, the reasoning system will determine where a concept can be located in a taxonomy hierarchy (a hierarchy built by means of a subconcept relation). Besides, the reasoner is used to check the consistency of the formal ontologies. Here, the validity of intentional definitions (in TBox) is checked. If an inconsistency is found, an expert user is responsible for solving it.

As result for each system, a normalized ontology (that can be divided into a top-level and a domain ontology) is returned. This ontology will be based on the geographic standards containing metaclasses descriptions (GFM) and the geographic schemas on which they are based. Thus, after passing through the logic process, the ontologies will have the correct structure we need to start with the following phase.

In the *Integration Phase* three processes are responsible for matching two normalized ontologies in order to create the global ontology. It contains the general concepts users will use to query the integrated system. In addition, a set of mappings are returned in order to represent the matching among the ontologies.

Merge, *General Analysis*, and *Specialized Analysis* are the processes of this phase. To do the first process, both ontologies of each system are joined by using generalization / specialization relations. In this way, the ontologies are taken as they are returned from the unit phase. Then, the two ontologies belonging to two different systems are merged. The merge process is performed by matching the classes that are part of the standard (metaclasses). As both ontologies have the same superclasses, merging is an easy task.

Once the merge process is finished, the *General Analysis* starts. It applies two types of analysis: syntactic and semantic. Within the syntactic analysis three syntactic functions are used in order to compare the names of the concepts in a

different way. Thus, functions return a different similarity result depending on the syntaxes of the compared names.

Then, in the semantic analysis, a thesaurus is used to extract synonym relationships between the concepts of the ontologies. These relationships are necessary because synonyms (in general) are not similar syntactically. In this case, WordNet[6] is used as the thesaurus. The *Specialized Analysis* performs a structural comparison by applying the similarity function described in [6, 14]. This function compares the number of properties that the classes have in common and analyzes them in a hierarchy (by calculating the depth of the most common superclass between the classes).

In the two last processes, user interaction is needed in order to determine the correct mapping. In this way, processes are user-driven and users are responsible for the final decisions.

Finally, it is possible the processes executed before generate inconsistencies within this final ontology. Therefore, the *System Phase* re-normalizes the global ontology created in the last phase. Like in the unit phase, a logic process is applied, where the reasoning system is used once more to analyze possible subsumption relations and inconsistencies in the global ontology.

User participation is also needed in this phase. Users here have two types of responsibilities – committing the options the reasoner system detects and testing the global ontology.

4 Related Work

Mapping discovery by using ontologies has being extensively investigated during the last years. Various approaches have emerged proposing processes and techniques to find similarities between elements of different but related ontologies.

In particular we are interested in methods for integration of geographic sources. In general, we can find three main overlapped mechanisms to perform integration, *the use of top-level ontologies*, *logical inferences* and/or *matching functions*. Table 1 shows the more representative and referenced proposals classified by these three types.

One particularity of all these proposals is the use of ontologies to represent either top-level information or domain information or both of them. In the case of ODGIS several ontologies are built (top-level, domain, and application ontologies) in order to provide more information about the domain and thus facilitate the integration process. But the activity of creating these ontologies is not an easy task and it demands a lot of effort. Other proposals as GeoNis, Aerts et al. and Hakimpour et al. use a top-level ontology together with the advantages of a formal language (to make inferences) as tools to find more suitable mappings. The use of similarity functions, in proposals as MDSM and SIM-DL, involves a set of functions that analyze the concepts and properties syntactically and semantically. In particular the use of these types of functions is useful when

[6] `http://wordnet.princeton.edu/`

Table 1. The three mechanisms for integration mapped to the proposals

	Top-level ontology	**Logical Inferences**	**Matching Functions**
BUSTER [15]		√	
Hakimpour et al. [16]	√	√	
MDSM [14]			√
ODGIS [17]	√		
GeoNis [18]	√	√	
Aerts et al. [19]	√	√	
Buccella et al. [5]	√	√	√
SIM-DL [20]			√

the ontologies are not complete (that is, there is absent information about the domain) and/or as starting point of an integration process when a top-level ontology is not involved. Proposals performing some manual step within the integration process require the assistance of an expert user to do so. For example, BUSTER needs of an expert user although it uses inferences during the query process.

Our merging method applies the three mechanisms to integrate ontologies. On one hand, top-level ontologies are created by using the information provided by the geographic standard. Then, logic capabilities and matching functions are combined in order to find more suitable mappings. The use of these three options makes our approach take advantage of the inherent benefits of using the standard in geographic information, the logic of data, and the semantic information from ontologies.

5 Conclusion and Future Work

In this work, we have presented an extension of the current Information Broker System in order to add capabilities which improve the generation of the federated schema. Particularly, our proposal aims at improving interoperability and consistency through the use of ontologies. However, there are still many issues that need further research. For example information sources in APAT Information Broker are not currently standadized, which may hinder consistency. The use of the ISO 19100 Stds. is a starting point for improving that. In addition, further validation of the ontology merging process would be absolutely neccessary for large ontologies – although our experiences [5] have shown good results when using small ones.

References

1. Calefato, F., Colagrossi, A., Gendarmi, D., Lanubile, F., Semeraro, G.: An information broker for integrating heterogeneous hydrologic data sources: A web services

approach. In Xu, A., Chaudhry, L., Guarino, S., eds.: Research and Practical Issues of Enterprise Information System, IFIP Series (Springer). Volume 205. (2006)
2. Gendarmi, D., Lanubile, F., Lichelli, O., Semeraro, G., Colagrossi, A.: Water protection information management by syntactic and semantic interoperability of heterogeneous repositories. In: Proceedings of the ISESS'07. (2007)
3. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2) (1993) 199–220
4. Baader, F., Calvanese, D., McGuiness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook - Theory, Implementation and Applications. Cambridge University Press (2003)
5. Buccella, A., Cechich, A.: Towards integration of geographic information systems. Electronic Notes in Theoretical Computer Science **168** (2007) 45–59
6. Buccella, A., Cechich, A., Brisaboa, N.R.: A three-level approach to ontology merging. In: MICAI'05. LNCS 3789, Monterrey, México, Springer-Verlag (November 2005) 80–89
7. : Geographic information. Rules for Application Schema. Draft International Standard 19109, ISO/IEC (2005)
8. : Geographic information. Geographic Information and Methodology for Feature Cataloguing. Draft International standard 19110, ISO/IEC (2005)
9. : Geographic information. Spatial Schema. International standard 19107, ISO/IEC (2003)
10. Belussi, A., Negri, M., Pelagatti, G.: An iso tc 211 conformant approach to model spatial integrity constraints in the conceptual design of geographical databases. In: ER (Workshops). (2006) 100–109
11. Jang, S., Kim, T.J.: Modeling an interoperable multimodal travel guide system using the iso 19100 series of international standards. In: Proceedings of the GIS '06, New York, NY, USA, ACM Press (2006) 115–122
12. Smith, M.K., Welty, C., McGuinness, D.: Owl web ontology language guide. W3C (February 2004)
13. Haarslev, V., Moller, R.: Racer system description. In Lambrix, P., Borgida, A., Lenzerini, M., Moller, R., Patel-Schneider, P., eds.: Proceedings of the CEUR-WS. Number 22, Linkoeping, Sweden (August 1999)
14. Rodríguez, M., Egenhofer, M.: Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. International Journal of Geographical Information Science **18**(3) (2004) 229–256
15. Visser, U.: Intelligent Information Integration for the Semantic Web. Volume 3159 of Lecture Notes in Computer Science. Springer Berlin - Heidelberg (2004)
16. Hakimpour, F.: Using Ontologies to Resolve Semantic Heterogeneity for Integrating Spatial Database Schemata. PhD thesis, Zurich University (2003)
17. Fonseca, F.: Ontology-driven Geographic Information Systems. PhD thesis, University of Maine (2001)
18. Stoimenov, L., Stanimirovic, A., Djordjevic-Kajan, S.: Discovering mappings between ontologies in semantic integration process. In: Proceedings of the AGILE'06, Visegr, Hungary (2006) 213–219
19. Aerts, K., Maesen, K., van Rompaey, A.: A practical example of semantic interoperability of large-scale topographic databases using semantic web technologies. In: Proceedings of the AGILE'06, Visegr, Hungary (2006) 35–42
20. Janowicz, K.: Sim-dl: Towards a semantic similarity measurement theory for the description logic *cnr* in geographic information retrieval. In: OTM Workshops (2). (2006) 1681–1692